ACOUSTIC SCENE CLASSIFICATION USING ENSEMBLES OF DEEP RESIDUAL NETWORKS AND SPECTROGRAM DECOMPOSITIONS

Technical Report

Yingzi Liu[1], *Shengwang Jiang*[2], *Chuang Shi*[3], *Huiyong Li*[4]

University of Electronic Science and Technology of China, Chengdu, China
[1] Yingziliu @std.uestc.edu.cn
[2] swjiang @uestc.edu.cn
[3] shichuang @uestc.edu.cn
[4] hyli@uestc.edu.cn

ABSTRACT

This technical report describes ensembles of convolutional neural networks (CNNs) for the task 1 / subtask B of the DACSE 2020 challenge, with emphasis on the use of a deep residual network applied to different spectrogram decompositions. The harmonic percussive source separation (HPSS), nearest neighbor filter (NNF), vocal separation and Head-related transfer function (HRTF) are used to augment the acoustic features. Our system achieves higher classification accuracies and lower log loss in the development dataset than baseline system.

***Index Terms*—** DCASE 2020, acoustic scene classification, deep residual network, spectrogram decomposition, HRTF

1. INTRODUCTION

How to make the machine accurately perceive and understand the content of acoustic scene like the human beings has always been a research hotspot in the field of audio signal processing. Acoustic scene classification (ASC) is an important technique for natural acoustic scene calculation and analysis. It refers to the task of associating a semantic label to an audio stream that identifies the environment in which it has been produced. ASC can be used in the design of context-aware services, intelligent wearable devices, robotics navigation systems, and audio archive management [1]. For example, the smartphones can continuously sense their surroundings by ASC and switch to silent mode when we enters a noisy place.

The DCASE Challenge (Detection and Classification of Acoustic Scenes and Events) is a technical competition sponsored by the audio and acoustic signal processing (AASP) technical committee, IEEE signal processing society (SPS). It is one of the most authoritative international evaluation and competition in the field of audio signal processing and focuses on Acoustic scene Classification, Acoustic event Detection and identification. Acoustic scene classification is a regular task in the DCASE challenge series, being present in each of its editions up until now [2]. In 2018 and 2019, the task of acoustic scene classification has been divided into three subtasks. The subtask B focus on classifying acoustic scenes with mismatched recording devices. It means that some devices appear only in the evaluation dataset. In 2020, the subtask A of acoustic scene classification is an extension of 2019.There are not only 4 real devices but also 11 simulated devices in this subtask which focus on generalization properties of systems across a number of different devices.

Each consecutive edition of the challenge has brought a new and larger dataset than previous edition, making it possible to use deep neural networks that rely on large amounts of data for training. Especially in 2018, the expansion of dataset promotes the use of convolutional neural network [3]. In the last two years of challenges, most of the top is a method using the convolutional neural network architecture.

Past entries into DCASE challenges have used the spectrogram and its variants for CNN processing, such as the short-time Fourier transform (STFT), log mel spectrogram, mel frequency cepstral coefficients (MFCC), constant-Q transform (CQT) [4]. In DCASE 2019 challenge, most of participants chose the log mel spectrogram as the features in their system. So we prefer to extract log mel spectrogram in our system. Moreover, spectrogram decomposition and HRTF are introduced to augment the features. By using these features for training, we can obtain several CNN models for ensemble. Past results have shown that ensemble can greatly improve system performance [5].

Figure 1 shows the architecture of the proposed system. The log mel spectrogram is divided by four dividing methods, including HPSS, NNF, HRTF and vocal separation. Thus we can obtain eight different features to train CNN models and finally use ensemble learning to make final decision.

2. ARCHITECTURE

2.1. Network Architecture

We use the model architecture introduced by McDonnell et al [6], a CNN in which the frequency and time axes are treated differently (see Figure 2). It used two pathways in the residual network: one for high frequencies and one for low frequencies, that were fused just two convolutional layers prior to the network output. The architecture is tuned to achieve the good performance in DCASE2019 challenge. The input has 128 frequency dimensions, then these dimensions are split into two pathway.

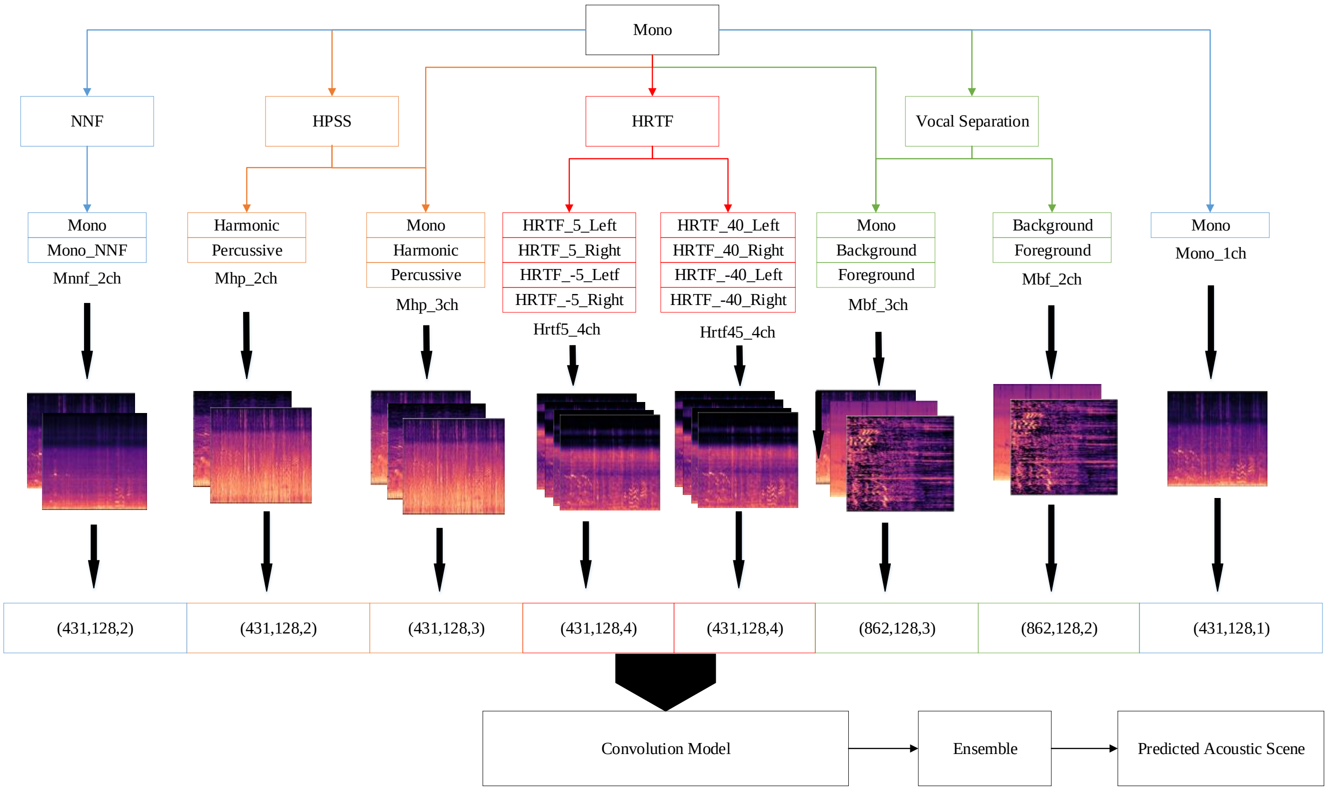


Figure1: Overall architecture. A monaural sample in the dataset is processed by the HPSS, NNF, HRTF and vocal separation. Eight different features are thus extracted to train CNN models. Ensemble of those CNN models provides the final decision [8]

The dimensions 0 to 63 is processed by a residual network with 17 convolutional and dimensions 64 to 127 by another. The kernels size in the two paths are 3 ×3. After these layers, we concatenate the two pathway separately to form 128 frequency dimensions, and then processed with two 1×1 convolutional layers. The second of these layers reduces to 10 classes. This is operated by a batch normalization layer, a global average pooling layer, and an activation layer used the softmax function. Additionally, when the number of channels needs to be increased before summation of different paths in the residual paths, zero padding only in the channel dimension is used. This networks had approximately 3.2 million trainable parameters.

2.2. Spectrogram Decomposition

2.2.1. Harmonic percussive source separation (HPSS)

Real world signals are usually mixtures of harmonic and percussive sounds. The goal of harmonic-percussive source separation (HPSS) is to decompose an audio signal into harmonic and percussive components. Applications of HPSS include audio remixing, improving the quality of chroma features, tempo estimation, and time-scale modification. Another use of HPSS is as a parallel representation when creating a late fusion deep learning system [7]. Many of the top performing systems of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 and 2018 challenges used HPSS for this reason. We use the Librosa package to implement the HPSS.

2.2.2. Nearest neighbor filter (NNF)

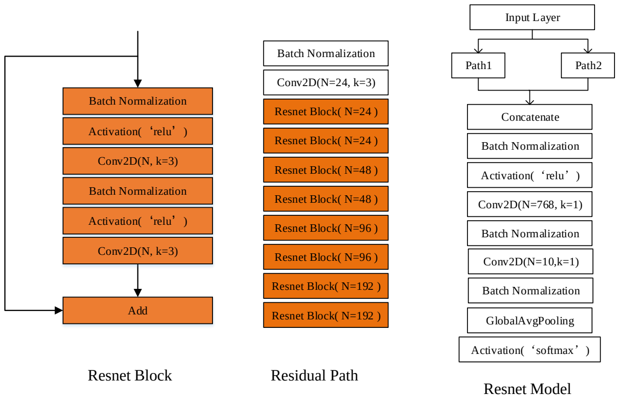
The NNF smooths the features to focus more on the overall picture instead of the details by removing the outliers. Using the NNF has achieved good performance in the past task of acoustic scene classification [8]. The Librosa package provides two options to do the NNF. One option is with non-local means method by setting ‘aggregate’ to ‘np.average’. The other option sets ‘aggregate’ to ‘np.median’. We choose the latter one in this technical report.

2.2.3. Vocal separation

The vocal separation can separate the vocal sound and background sound from the mixed audio data [9]. We think the background sound contains more information about the acoustic scene in the task of acoustic scene classification. So we try to use the vocal separation to separate the sporadic foreground signal from the background signal with certain patterns. Similarly, we use the Librosa package to implement the vocal separation with default settings.

2.3. HRTF

The HRTF is a transfer function that models the process how human's ears receive sound from a point in space [10]. Since everyone has two ears, the HRTF always appears in a pair. There are 25 azimuths (ranging from -80 to 80 degrees) and 50 elevations (ranging from -45 to 230.625 degrees) in the CIPIC database that we can use [11]. We select 24 azimuths and set the elevations to 0. Every two azimuths being symmetric to the median plane are paired together to provide a 4-channel feature [12].



Resnet Block

Residual Path

Resnet Model

Figure 2: Resnet Block, Residual Path and Resnet Model. N represents the number of output channels, and k represents the size of the convolution kernel

2.4. Features

The samples in the DCASE2020 task 1 / subtask A dataset are monaural and have a common sampling rate of 44.1 kHz. Preprocessing is done similar to [8]: First, we extract the spectrogram using a Short Time Fourier Transform (STFT) with a hamming window size of 2048 and 50% overlap. Then we apply the log mel filter bank on the spectrogram to get the log mel spectrogram. There are 128 log mel filters in the filter bank that cover a frequency range from 0 to 22.05 kHz, yielding 431-frame spectrograms with 128 frequency bins. The log mel spectrograms are normalized by subtracting the mean and dividing the standard deviation.

Therefore, by combining the output of the NNF with the input, we obtain the (431, 128, 2) feature 'Mnnf_2ch'. The output of the HPSS results in a two-channel feature 'Mhp_2ch' with the size of (431, 128, 2) and together with the input forms a three channel feature 'Mhp_3ch'with the size of (431, 128, 3). Similarly, the vocal separation also provides the (862, 128, 3) feature 'Mbf_3ch' and the (862, 128, 2) feature 'Mbf_2ch' with and without the input, respectively. Note that the hop size setting in the vocal separation is different from that of the others. This results in a change of the feature size. Moreover, we obtain the 'Hrtf_4ch' features by using the HRTF. In our system we use two (amazimuth=5 and azimuth=40) and all HRTF features to train CNN models respectively

2.5. Network Ensemble

Ensemble learning has placed first in many prestigious machine learning competitions and it helps improve machine learning results by combining several machine learning techniques into one predictive model. This approach can achieve higher classification accuracy and better generalization compared to a single model [12]. In order to exploit independence between the base learners and reduce the error dramatically by averaging, we train some CNN models in parallel and then put them together to make the final decision.

There are many ensemble strategies that we can choose. In this technical report, we use averaging and stacking for comparison. The averaging method averages the output probabilities of different models, as shown in Figure 3. The stacking combines multiple classification models via a meta-classifier. The base level learners are trained based on a complete training set, then the meta-learner is trained on the outputs of the base level model as features. We choose the CNNs to be the base learners and the random forest to be the meta-learner, as illustrated in Figure 4.

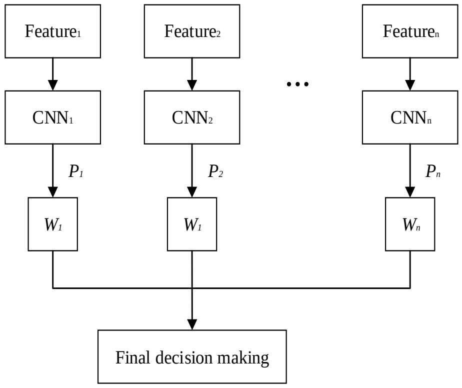


Figure 3: Weighted averaging of the CNN ensemble [8]

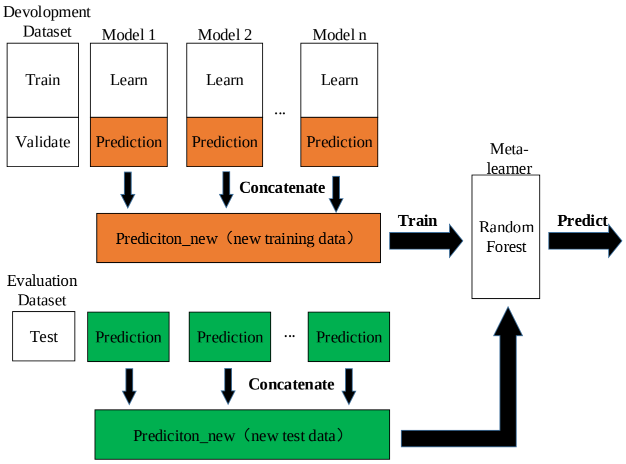


Figure 4：Stacking of CNNs by the random forest [8]

3. EXPERIMENTS

3.1. Dataset

Similarly to 2019 DCASE challenge, the dataset of the task 1/subtask A of the DCASE 2020 challenge contains recordings in 10 different acoustic scenes using 4 real devices(referred to as A, B, C and D). Additionally, in order to simulate realistic recordings, synthetic data for 11 mobile devices S1-S11 is created based on the audio recorded with device A. The development set contains data from 10 cities and 9 devices: 3 real devices (referred to as A, B, C) and 6 simulated devices (referred to as S1-S6). The total amount of audio in the development set is 64 hours. The evaluation dataset contains 33 hours of audio from 12 cities and 11 devices, including 1 real device D and 5 simulated devices S7-S11 that are not present in the development set.

3.2. Results and Submissions

Models are trained using an SGD optimizer with a batch size of 32, momentum of 0.9, decay of 0.0001, and the focal loss function. Each model is trained for 200 epochs. The initial learning rate is set to 0.01 and decreased by a factor 0.5 every 10 epochs after 50 epochs. Then, the model with the highest testing accuracy is saved. In the random forest, the number of decision trees are set to 5000. We also employ mixup and temporal crop augmentation during training, using the same approach as [6].

Table 1 lists the models that we submit. The main metric for this task is the macro-average accuracy (average of the class-wise accuracies).In order to have a metric which is independent of the operating point, the multiclass cross-entropy (Log loss) is used as a secondary metric. All submissions achieve higher classification accuracies in the development dataset than baseline system. The last two submission used random forest ensemble strategy are get lower log loss than baseline.

Table 1: Results of development dataset

Method	Accuracy	Log loss
Baseline	0.541	1.365
Averaging_8	0.684	1.362
Averaging_18	0.690	1.367
Randomforest_8	0.687	0.841
Randomforest_18	0.684	0.839

- **Averaging_8** is the averaging ensemble of 8 models that used only two feature by HRTF.
- **Averaging_18** is the averaging ensemble of 16 models that used all feature by HRTF.
- **Randomforest_8** is the stacking ensemble of 8 models that used only two feature by HRTF.
- **Randomforest_18** is the stacking ensemble of 16 models that used all feature by HRTF.

4. CONCLUTIONS

In this technical report, we have described a system for the task 1/subtask A of the DCASE 2020 challenge. We use NNF, HPSS, vocal separation, and HRTF to augment the acoustic features. Our system used a deep residual network with two path as a base learner for the ensemble. The averaging and stacking are used as ensemble strategies. Additionally, the mixup and temporal crop augmentation also boost performance on all devices. The experiment results over DCASE2020 development dataset targeting task 1A review that our method are effective to improve the classification accuracy over every class.

5. REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic scene classification: Classifying environments from the sounds they produce,” *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] http://dcase.community.

[3] A. Mesaros, T. Heittola, and T. Virtanen, “A multi-device dataset for urban acoustic scene classification,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, Surrey, UK, Nov.2018.

[4] Michał Kosmider, “Calibrating neural network for secondary recording devices,” in *the 2019 Detection and Classification of Acoustic Senes and Events Challenge,* Warsaw, Poland, July.2019.

[5] Khaled Koutini1, Hamid Eghbal-zadeh and Gerhard Widmer “CP-JKU submission to Dcase’2019: Acoustic scene classification and audio tagging with receptive-field-regularized CNNs,” in the *2019 Detection and Classification of Acoustic Scenes and Events,* Austria, July. 2019.

[6] Mark D. McDonnell and Wei Gao, “Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths,” in *the 2019 Detection and Classification of Acoustic Senes and Events Challenge,* Australia, July.2019.

[7] Driedger, J, M. Muller, and S. Disch, “Extending harmonic-percussive separation of audio signals,” *Proceedings of the International Society for Music Information Retrieval Con*ference. Vol. 15, 2014.

[8] Shengwang Jiang, Chuang Shi and Huiyong Li, ”Acoustic scene classification using ensembles of convolutional neural network and spectrogram decompositions,” in the *2019 Detection and Classification of Acoustic Scenes and Events,* Chengdu, China, July. 2019.

[9] Z. Rafii and B. Rardo, “Music/Voice separation using the similarity matrix,” in *the International Society for Music Information Retrieval Conference,* Porto, Portugal, Oct. 2012.

[10] Y. Iwaya, M. Otani, T. Tsuchiya and J. Li, “Virtual auditory display on a smartphone for high-resolution acoustic space by remote rendering,” in *the 2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing,* Adelaide, Australia, Sep. 2015.

[11] V. Algazi, R. Duda, D. Thompson and C. Avendano, “The CIPIC HRTF database,” in *the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics,* New Paltz, NY, Oct. 2001.

[12] Shengwang Jiang, Chuang Shi and Huiyong Li, ”Acoustic scene classification technique for Active noise Contral,” in the 8th International Conference on Control, Automation, Chengdu, China, Oct. 2019